Structure
Ways & Means

Systematic Bioinformatics and Experimental Validation of Yeast Complexes Reduces the Rate of Attrition during Structural Investigations

Mark A. Brooks,[1,9,10] Kamil Gewartowski,[2,10] Eirini Mitsiki,[3,10] Juliette Létoquart,[4] Roland A. Pache,[5] Ysaline Billier,[4] Michela Bertero,[6] Margot Corréa,[4] Mariusz Czarnocki-Cieciura,[3] Michal Dadlez,[2] Véronique Henriot,[4] Noureddine Lazar,[1] Lila Delbos,[1] Dorothée Lebert,[4] Jan Piwowarski,[2] Pascal Rochaix,[4] Bettina Böttcher,[7] Luis Serrano,[6,8] Bertrand Séraphin,[4,11,*] Herman van Tilbeurgh,[1,*] Patrick Aloy,[5,8,*] Anastassis Perrakis,[3,*] and Andrzej Dziembowski[2,*]
[1]IBBMC-CNRS UMR8619, IFR 115, Bât. 430, Université Paris-Sud, 91405 Orsay, France
[2]Department of Genetics and Biotechnology, Warsaw University and Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawińskiego 5a, 02106 Warsaw, Poland
[3]Department of Biochemistry, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[4]Equipe Labelisée La Ligue, CGM, CNRS UPR2167, Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France
[5]Institute for Research in Biomedicine (IRB) and Barcelona Supercomputing Center (BSC), c/ Baldiri I Reixac 10-12, 08028 Barcelona, Spain
[6]Systems Biology Laboratory, Centre for Genomic Regulation, Barcelona 08003, Spain
[7]School of Biological Sciences, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, UK
[8]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
[9]Present address: Evotec (UK) Ltd., 114 Milton Park, Abingdon, Oxon, OX14 4SA UK
[10]These authors contributed equally to this work
[11]Present address: IGBMC, 1 rue Laurent Fries, BP10142, 67404 Illkirch, France
*Correspondence: seraphin@igbmc.fr (B.S.), Herman.Van-Tilbeurgh@u-psud.fr (H.v.T.), patrick.aloy@irbbarcelona.org (P.A.), a.perrakis@nki.nl (A.P.), andrzejd@ibb.waw.pl (A.D.)
DOI 10.1016/j.str.2010.08.001

SUMMARY

For high-throughput structural studies of protein complexes of composition inferred from proteomics data, it is crucial that candidate complexes are selected accurately. Herein, we exemplify a procedure that combines a bioinformatics tool for complex selection with in vivo validation, to deliver structural results in a medium-throughout manner. We have selected a set of 20 yeast complexes, which were predicted to be feasible by either an automated bioinformatics algorithm, by manual inspection of primary data, or by literature searches. These complexes were validated with two straightforward and efficient biochemical assays, and heterologous expression technologies of complex components were then used to produce the complexes to assess their feasibility experimentally. Approximately one-half of the selected complexes were useful for structural studies, and we detail one particular success story. Our results underscore the importance of accurate target selection and validation in avoiding transient, unstable, or simply nonexistent complexes from the outset.

INTRODUCTION

Numerous large-scale proteomics initiatives in the model organism *Saccharomyces cerevisiae* have been reported over the last few years and have provided evidence for thousands of new protein interactions and supplied a wealth of information about the composition of macromolecular complexes (Gavin et al., 2006; Ho et al., 2002; Ito et al., 2001; Krogan et al., 2006; Tarassov et al., 2008; Uetz et al., 2000). Nevertheless, the characteristics of protein interaction networks in vivo have not yet been rigorously untangled for any organism, let alone the faithful budding yeast. Now that such protein interaction data sets are in the public domain, a gauntlet has been thrown down to the scientific community to provide tools for assimilating these data with a view to developing algorithms and experimental methodologies for predicting the composition of complexes with high accuracy, thereby facilitating their functional and structural characterization.

However, for many predicted complexes identified in high-throughput affinity purification experiments, their subunit composition is not established with sufficient reliability to proceed to structure determination. Improvements in the confidence that can be placed in protein interaction models are therefore clearly needed, with the specific aim of identifying complexes with well-defined stoichiometry, and which are amenable to structural studies. Raising the confidence with which complex composition could be predicted would benefit enormously the field of structural biology. Ideally, it would be possible to identify stable complexes (for example, ribosomes, RNA polymerases, the exosome, or the 20S proteosome) and discriminate them from more dynamic assemblies that contain transient interactors (for example, spliceosomes or the 26S proteosome). It would therefore be beneficial to classify and characterize the various entities which form the central frameworks of protein-protein interaction networks (Gavin et al., 2006; Higurashi et al., 2008; Krogan et al., 2006).

Foremost among the problems encountered in complex characterization are those related to the primary data being of limited quality. For example, the heterogeneity or the extremely dilute nature of samples from proteomic experiments results in complex subunits being overlooked. Additionally, in some studies, the characterization of complex composition has been hindered by the contamination of bona fide complexes by so-called ''background'' or ''sticky'' polypeptides that interact with other proteins in a promiscuous fashion (Shevchenko et al., 2002). One challenge is therefore to devise a computational strategy to filter through the results of many thousands of biochemical purifications which have been performed to date, and identify the complexes that will yield the optimal results during expression and purification studies (Bravo and Aloy, 2006).

The first structural genomics consortia focused on the determination of X-ray and NMR structures at the level of the single protein (Alzari et al., 2006; Graslund et al., 2008; Marsden and Orengo, 2008). More recently, the Structural Genomics Consortium (SGC) (Edwards et al., 2002), the 3D Repertoire (http://www.3drepertoire.org/) and SPINE 2 - Complexes (http://www.spine2.eu/) consortia have opted to study macromolecular complexes from a medium-throughput perspective. The expression and purification of protein complexes adds an extra level of complexity, since globular protein interfaces are often partly hydrophobic, and single partners may be insoluble. In many cases, only in the context of an assembled complex do hydrophobic interfaces become buried and the participating polypeptides can be produced as soluble entities (Dyson and Wright, 2005; Smialowski et al., 2007).

Since the inception of the European Commission-funded consortium ''3D repertoire'' in 2004, collaborating scientists have been addressing the problems associated with identifying complexes de novo for structural studies. Within the first step, which consisted of highly selective filtering of existing data sets for evidence of the existence of complexes in a process we term ''complex triage,'' three methods were employed. First, a bioinformatics-based selection procedure, optimized using a training set composed of complexes of known three-dimensional (3D) structure, was used to screen for stable, well-folded complexes. Second, we examined the results from high throughput affinity purification experiments manually, focusing on the visual inspection of gels to identify complexes of which the components existed in stoichiometric quantities. Finally, a set of seven complexes was chosen on the basis of the scientific literature.

A compilation of these complexes, named the ''list of 20,'' were then validated by new affinity purifications of the natural complexes and their subunit compositions were confirmed using mass spectrometry. In addition, the solution sizes of these complexes were assessed by size exclusion chromatography. The subset of proteins that were shown to indeed participate in macromolecular assemblies as predicted and that were also believed to be tractable for structural studies was then cloned and expressed in *Escherichia coli.* Using various techniques, we aimed to obtain purified material suitable for structural analysis. We show the overall success in each of the steps of this procedure and present a detailed account of one example complex. The results from this test set of complexes under investigation have allowed us to evaluate the effectiveness of each of the techniques used and devise an optimal route for the production of protein complexes in structural biology pipelines.

RESULTS

Identification of Complexes for Structural Studies

Complex Triage by Bioinformatics

A system based on the notion that complexes likely amenable to structural studies should be small, compact, and homogeneous has been previously described (Gavin et al., 2006). We considered biophysical, biochemical, and large-scale proteomics data in the form of partial scoring functions that were normalized and combined into a final feasibility score for each complex. The algorithms that define our scoring function, hereafter called the Complex Feasibility (CF) algorithm, have been published elsewhere (Pache and Aloy, 2008); details are available in http://gatealoy.pcb.ub.es/targetselection/help.html and the main issues are summarized in Supplemental Experimental Procedures (available online).

For the evaluation set of complexes we use in this study (Table 1; Table S3), we combined four of the top-ranking complexes made by the CF tool (which we expected to behave very well for structural studies), with three mid-ranking complexes (which we expected to present more of a challenge for structural studies).

Complex Triage by Manual Visualization of Gels

Complexes with apparently stoichiometric components are more likely to indicate stable interactions and be more suitable for structural studies. In the original genome-wide approach (Gavin et al., 2006), tandem affinity-purified (TAP) assemblies were separated by denaturing gel electrophoresis and stained. The gels were then cut into 1 mm slices, digested with trypsin, and analyzed by MALDI-TOF mass spectrometry (MS). This procedure did not take into account the relative quantities of proteins present in the TAP eluate.

We thus decided to visually inspect the original gels (Gavin et al., 2006) for bands indicative of stoichiometric complexes. Thorough inspection of about 4000 purification experiments identified 64 promising complexes (Table S4: dimeric complexes; Table S5: trimeric complexes; Table S6: tetrameric complexes). Some of the 64 chosen complexes have not been identified as being complexes in the original automated annotation (Gavin et al., 2006) and were thus not considered in the CF classification algorithm. Notably, the best six complexes that were chosen independently by gel inspection were all in the top 50 of the CF algorithm, and two of them were in the top 10. Six complexes were finally selected by manual gel inspection (Table 1).

The List of ''20 Complexes''

We finally selected 20 complexes; the corresponding bioinformatics and gel scores, and when possible appropriate references to the literature are summarized in Table 1. Although the manual gel inspection and the bioinformatics efforts were independent, all previously identified complexes selected by manual screening had a high ranking using the CF algorithm. In contrast, not all of the complexes chosen by the algorithm could be associated with clear and conclusive gels. Notably, a top-ranked choice was associated with a gel of mediocre quality. Nonetheless, such types of selections resulted in a potentially interesting collection of complexes that

Table 1. Summary of Target Selection, Validation, and Complex Reconstitution Results

Complex \ Stage	Selection		Validation	Single subunit expression		Complex production		
	Gel Quality	Rank	TAP Results	Subunit 1	Subunit 2	Reconstitution[a]	Separate plasmid co-expression	Operon co-expression
Bioinformatics Analysis								
Ste11, Ste50	Good	1	Heterogeneous	+	-	-	ND	ND
Atg17, Atg20, Atg29	Good	27	Partial (-Atg29)	+	-	ND	-	-
Vps27, Hse1	Excellent	1	Excellent	+	+	+	+	ND
Psy2, Psy4, Pph3	Excellent	4	Partial (-Pph3)	+	-	ND	-	ND
Nup82, Nup159, Nsp1	Good	4	Excellent	ND	ND	ND	ND	ND
Ede1,Syp1	Good	22	Excellent	ND	ND	ND	ND	ND
Dop1, Mon2	Excellent	25	Aggregated	ND	ND	ND	ND	ND
Gel Analysis								
Gcd14, Gcd10	Excellent	12	Excellent	+	+	+	+	+
Ptc2, Paa1	Excellent	8	Paa1 promiscuous	+	+	+	ND	ND
Met12, Met13	Excellent	22	Excellent	+	+	ND	+	ND
Dug3, Dug2	Excellent	9	Excellent	+	+	ND	+	ND
Ssl2, Yor352w	Excellent	27	Excellent	+	+	ND	+	+
Spt6, Spn1	Excellent	40	Partial (-Spn1)	ND	ND	ND	ND	ND
Literature Analysis								
Rad17, Mec3, Dcd1	Failed	261	Failed	ND	ND	ND	ND	ND
Orc1-6	Good	29	Heterogeneous	ND	ND	ND	ND	ND
Rbg2, Gir2	Good	5	Excellent	+	+	ND	+	+
Dom34, Hbs1[b]	No Interaction	364	No Interaction	+	+	+	ND	ND
Rps28B, Edc3	No Interaction	364	Edc3 promiscuous	ND	ND	ND	+	+
Sis2, Ykl088w, Vhs3	Partial Interaction	323	Weak expression	ND	ND	ND	ND	ND
Mtw1,Dsn1,Nnf1,Nsl1	Partial Interaction	11	Weak expression	ND	ND	ND	ND	ND

Complexes selected by bioinformatics, gel, and literature analyses, respectively, are listed. The complexes were assessed according to their purity after TAP purification (column labeled "Gel quality"). The ranks according to the CF algorithm of each of the complexes ("Rank"), as well as the results of validation by tandem affinity purification (cf. Figure S1; "TAP Validation" and Figure S2 for the results of complex production and Table S4) are shown. Results of expression, coexpression, and reconstitution studies are as follows: +; successful, -; unsuccessful, ND; not determined, NA; not applicable. See also Tables S1, S2, and S3.

[a] The Gcd10:Gcd14 complex was not reconstituted from purified proteins, but instead cells in which the proteins had been expressed separately were combined prior to sonication. For clarity, results that were deemed to be "positive" (having a "good" gel quality, high ranking in the bioinformatics triage, significant expression levels or production of the relevant complex by either coexpression or by reconstitution) are shown with a green background. Similarly, "mediocre" results in the TAP validation (indicating that either heterogenous or partial complexes were purified) are shown with a yellow background. Negative results, indicating either a poor gel quality, low bioinformatics rank, failed TAP validation experiment, failed expression, or failed complex production, are shown in red. Expression results for the complexes not deemed to be suitable for structural analysis are shown as gray text.

[b] Reconstitution of the Dom34:Hbs1 complex had been described previously (Graille et al., 2008).

would hopefully be amenable to structural studies. The selection was complemented by the choice of an additional seven complexes suggested by partners of 3D repertoire, based on specific biological interests and literature know-how, reaching the final number of 20 complexes included in this study. Interestingly, only one of the latter choices was in the top 10 bioinformatics list, and an additional two were in the top 50; the remaining four scored poorly in the CF algorithm.

Validation of Complex Composition

The 20 selected complexes were validated in a two-step TAP purification on IgG and calmodulin columns. Mass spectrometry analyses using an ESI-TRAP approach were performed using both the eluate solutions and the excised gel bands as samples. In addition, molecular weights of complexes were estimated by size exclusion chromatography of total extracts, followed by dot-blot detection of TAP-tagged proteins in eluate fractions. Finally, the molecular weights of tagged subunits and the efficiency of binding to IgG resin were verified by western blot analyses (see Figure 1 for a schematic representation of the procedure). The conclusions regarding individual complexes are presented in Table 1 and Figure S1.

Only two of the complexes completely failed this validation stage, one for technical reasons and one could not be identified

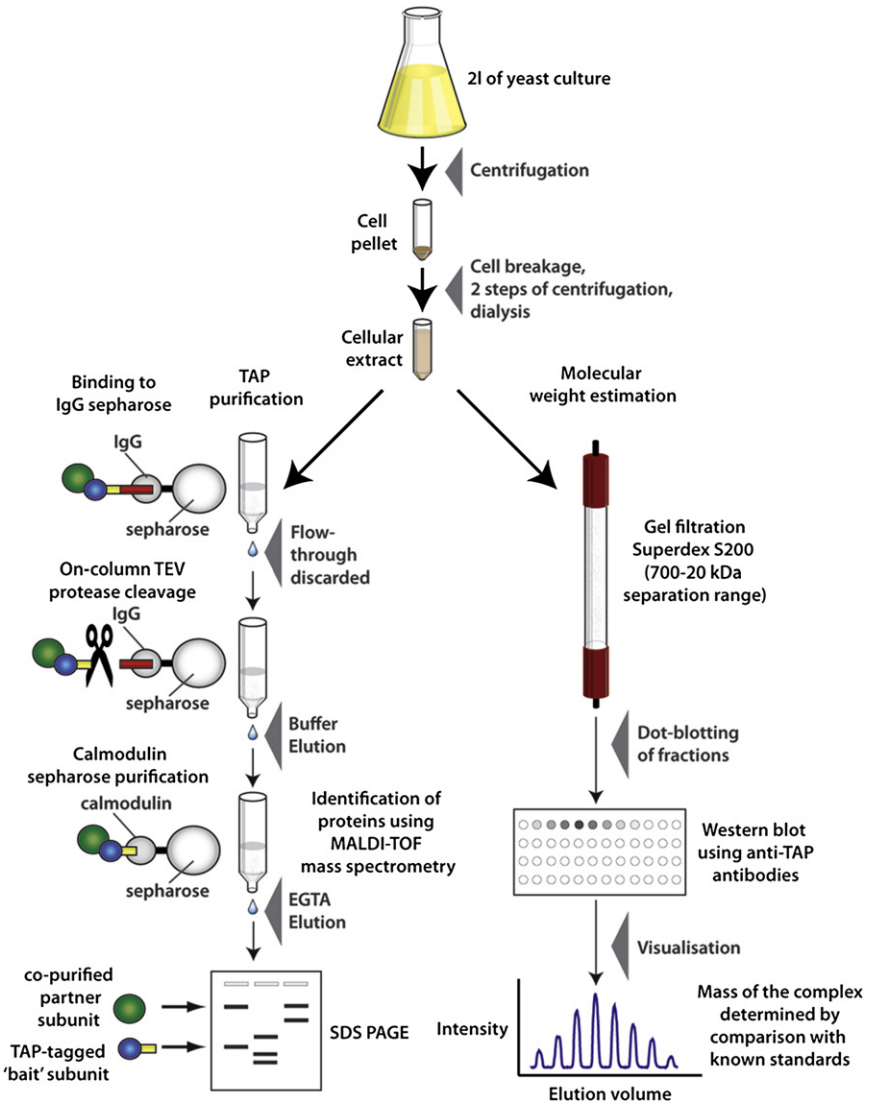

Figure 1. Strategy for the Validation of Selected Complexes

A schema showing the overall pathway for the validation of complex composition and estimation of molecular weight of each complex is presented. The complexes were expressed in yeast using a C-terminal TAP-tag of the bait protein. Following cell breakage, complexes were either subjected to TAP purification to assess the subunit composition, or to gel filtration in order to estimate the molecular weight, and thereby their stoichiometry. See Figure S1 for actual results of the validation experiments.

at all. Interestingly, both complexes originated from the literature additions to the list and they both scored poorly in the bioinformatics assessment. This category, where literature knowledge was used to select complexes, gave a lower validation rate than the other strategies. Apart from the one complex for which no technically valid results were obtained, one failed, while two others showed too weak native expression to be conclusive. Another complex was highly heterogeneous and one included a very promiscuous protein as a partner and was thus inconclusive. Notably, one complex selected from the literature and validated here to be "excellent" was ranked in the top 10 (20th percentile) of the bioinformatics list. The low validation rates of complexes selected from the literature, and their low biocomputing ranks stem from their specific characteristics (low abundance, specific interaction involving abundant partners flagged as promiscuous) and underline the limitation of current strategies to identify bona fide complexes. The gel-selected complexes and the bioinformatics complexes fared well in the validation, with four out of six and three out of seven, respectively, being scored as "excellent." From the validated complexes, 11 were chosen for heterologous expression studies and production in quantities suitable for structural studies. Analysis of the twelfth complex, Dom34:Hbs1 is described elsewhere, so was not repeated (Graille et al., 2008) but is included in Table 1.

Recombinant Production of Complexes for Structural Studies

For these 11 complexes, a mixture of expression strategies was employed for their evaluation: expression of the full-length individual subunits, in vitro complex reconstitution from subunits, and coexpression. A total of 22 proteins have been used in expression trials as single full-length proteins in *E. coli*, either from synthetic, codon-optimized genes (16 proteins, Figure 2A) or from natural yeast genes (Figures S2–S6). Only three of these failed to produce soluble protein in appreciable amounts (Atg29, Psy4, and Ste11). We proceeded to reconstitute four complexes (Vps27:Hse1, Ptc2:Paa1, Ste11:Ste50, and Gcd10:Gcd14) from individually purified partners and succeeded in purifying three in soluble form and defined subunit composition (Figures S3–S6). In parallel, we also attempted coexpression of nine complexes: in these experiments only one of the two subunits was N-terminally tagged with a 6-His tag. In seven out of nine cases, we were able to produce both proteins and purify them as a complex by metal affinity chromatography (Figure 2B).

A Case Study of an Example Complex, from Selection to Validation

To illustrate the course of an experiment from target selection to validation, we present one particular exemplary complex. The Gcd10:Gcd14 complex was originally identified a few years ago and purified as a dimeric tRNA(1-methyladenosine) methyltransferase (Anderson et al., 1998, 2000; Ozanick et al., 2007). Gavin et al. (2006) observed again this dimeric complex, which was annotated as Complex 376 in the Krogan et al. (2006)

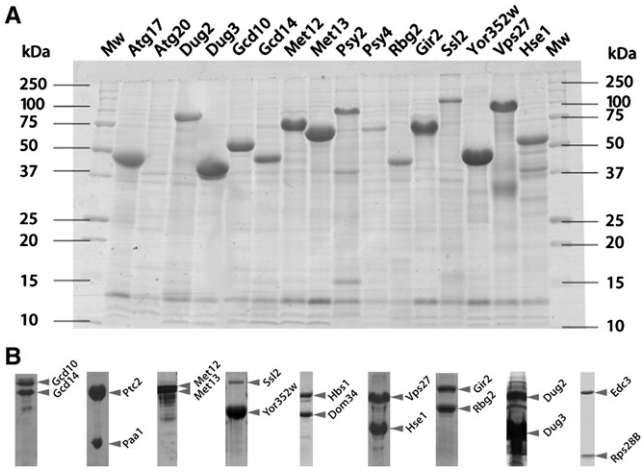

Figure 2. Expression and Purification of Yeast Full-Length Proteins

(A) SDS-PAGE analysis of full-length yeast constructs produced using codon-optimized synthetic genes, Ni^{2+}-NTA-purified and visualized using Coomassie. Full-length proteins were expressed and purified as above and eluted material was analyzed by SDS-PAGE. The samples are relatively pure after only one step of purification, although degradation products are sometimes present. Molecular weight markers and their sizes are indicated on both sides of the gel. Successful constructs are Atg17 (48.7 kDa), Dug2 (98.1 kDa), Dug3 (40.2 kDa), Gcd10 (54.4 kDa), Gcd14 (43.9 kDa), Met12 (73.9 kDa), Met13 (68.6 kDa), Psy2 (98.1 kDa), Rbg2 (41 kDa), Gir2 (31 kDa), Ssl2 (95.3 kDa), Yor352w (39.3 kDa), Vps27 (71.9 kDa), Hse1 (51.1 kDa), while the unsuccessful constructs are Atg20 (72.5 kDa) and Psy4 (50.7 kDa).

(B) The nine complexes successfully produced in a recombinant form. Ni^{2+}-NTA-purified samples of the results of complex formation trials were subjected to SDS-PAGE analysis and visualized using Coomassie. For coexpressed complexes the tagged component is marked with an asterisk below, while reconstituted complexes are not marked: Gcd10:Gcd14 (54.4 and 43.9 kDa, respectively), Paa1:Ptc2 (21.9 and 50.3 kDa), *Met12:Met13 (73.9 and 68.6 kDa), Dug2:*Dug3 (98 and 40.2 kDa), Ssl2:*Yor352w (95.2 and 40.2 kDa), Hbs1:Dom34 (68.7 and 44.1 kDa), Vps27:Hse1 (71.9 and 51.1 kDa), *Gir2:Rbg2 (31 and 41 kDa), Dug2:Dug3* (98.1 and 40.2 kDa), *Rps28B:Edc3 (7.6 and 61.3 kDa) complexes. Bands corresponding to the proteins of interest are arrowed.

See also Figures S2, S3, S4, S5, S6, and S7.

enumeration. TAP-purified Gcd10:Gcd14 has also been shown to be relatively homogeneous and therefore pure by electron microscopy. We selected this complex by gel analysis but it also ranked with a score of 12 by the CF algorithm.

First, we revalidated the complex by repeating the TAP purification using tagged Gcd14 and the only partner that was isolated was Gcd10, with no other bands either apparent or identified by mass spectrometry Figures 3A and 3B). Gel filtration analysis of the TAP-tag-purified complex was consistent with a molecular weight of approximately 350 kDa, suggesting the formation of higher order multimers since the expected mass of the Gcd10:Gcd14 complex with a 1:1 stoichiometry is 98.3 kDa.

The complex was reconstituted from the Ni^{2+}-NTA-purified individual components and subjected to gel filtration chromatography. The resulting complex had an approximate molecular weight of around 350 kDa, in agreement with the analysis of the "native" TAP-tagged complex (Figure 3B). The purified complex was then used in a negative stain electron microscopy experiment. The sample was homogeneous and could be used for data collection (Figure 3C). Image reconstructions without any imposed symmetry showed a tetrameric core with extensions at opposite surfaces, giving the entire complex 2-fold, as well as quasi 4-fold symmetry. Therefore, C2 symmetry was imposed for further refinement. The final reconstruction is shown in Figure 3E. Projections of this reconstruction agree with class averages were determined by multivariate statistical analysis (Figure 3D).

DISCUSSION

In this work, we set out to identify an optimal strategy for the analysis of *Saccharomyces cerevisiae* complexes by combining contemporary structural biology tools with the numerous proteome-level biochemical interaction data sets. Our central tenet was that we believed such data to be essentially reliable, and that the use of improved bioinformatics tools, manual analysis of gels or bibliographic curation of previous data should allow the identification of complexes best suited to structural analysis.

A question that we sought to answer related to whether bioinformatics, and specifically the CF algorithm, could provide trustworthy guidance when selecting targets. Ideally, the algorithm should eliminate the need for manual inspection of data. Therefore, we first generated a target list, partly using automated tools and partly manually. The next step was to ascertain which of the selected complexes do indeed exist in a stable and stoichiometric form. Our experimental results show that the bioinformatics algorithm could select targets with a validation success rate that was very high, and comparable to visual inspection of gels.

In the final CF algorithm, the most important parameters were the yeast two-hybrid ratio and the socio-affinity index (Table S2). The usefulness of the former parameter has been obvious for some time, since yeast two-hybrid screening has been a mainstay of research into protein-protein interactions. However, the important role of the socio-affinity index in this experiment was encouraging (Gavin et al., 2006), and we believe that it is a valuable and powerful metric for the identification of protein complexes based on protein interaction data sets. Conversely, the least useful parameters were the "average number of problematic residues" and the "colocalization ratio"; it appears that these parameters are not as useful as had been previously thought, at least in the context of this work (Pache and Aloy, 2008).

We note that some complexes identified by bibliographic analyses, which could not be validated and for which low scores were obtained with the CF algorithm, performed well using recombinant expression. These facts underline the limitation of complex analyses of low abundance complexes and/or complexes involving very abundant subunits for which it is difficult to exclude the existence of promiscuous interactions. It is possible that our complex triage procedures have been successful at least in part, due to the clarity of primary data for which the subunits are stoichiometrically equivalent and well expressed.

The success rate of obtaining soluble subunits by heterologous recombinant expression, for the full-length proteins was high (only 3 of 22 proteins tested could not be produced in

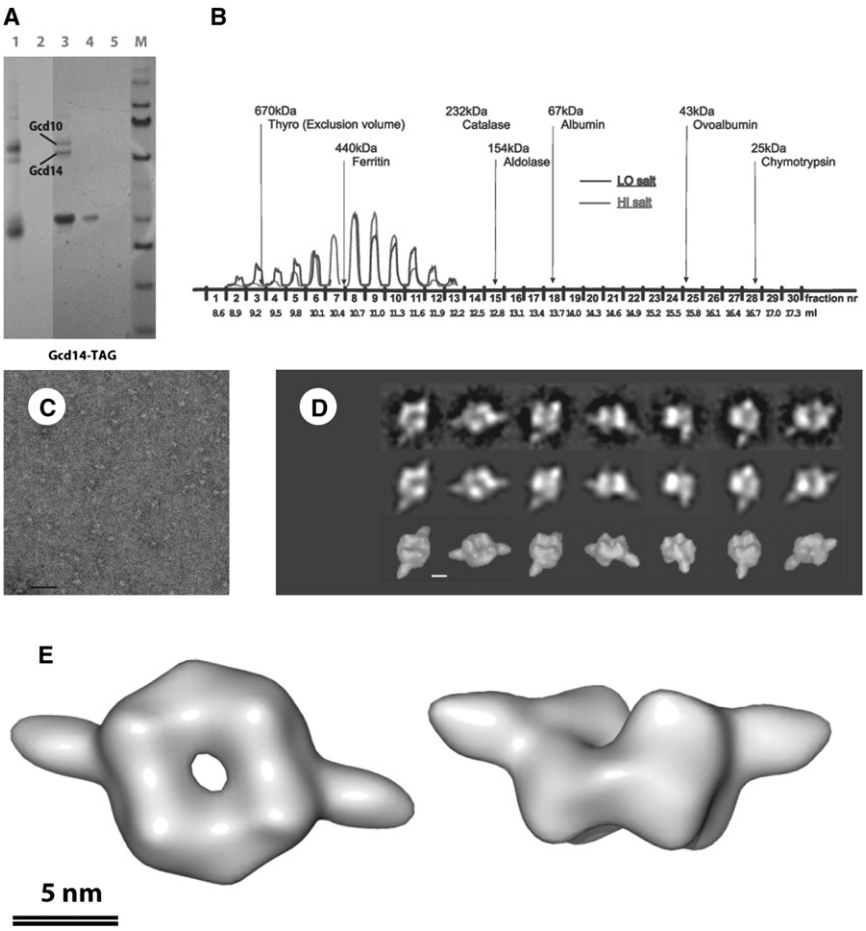

Figure 3. Validation and Scale-Up of an Exemplary Complex; Gcd10:Gcd14

(A) Both Gcd10 and Gcd14 were clearly visible after purification using the TAP protocol, with little evidence of contaminating proteins, validating this complex.

(B) In order to estimate the size of the complexes, yeast extracts were separated with the use of size exclusion chromatography on a Superdex 200 column in low (150 mM; marked as "LO") and high (500 mM; "HI") concentration of NaCl. Thirty fractions from this chromatography step were collected and spotted on a nitrocellulose membrane. To detect fractions containing the TAP-tagged protein, western blotting using PAP antibodies was performed. See the legend to Figure S1 for further details for (A) and (B).

(C) Micrograph of the Gcd10:Gcd14 complex which had been purified as in Figure 2B and fixed with glutaraldehyde, according to the GraFix protocols (Kästner et al., 2008) and stained with uranyl-acetate in a sandwich between two layers of carbon. Scale bar, 50 nm.

(D) Class averages of the data (top row) determined by multistatistical analysis agree with projections of the 3D map (central row). Surface presentations (bottom row) of the 3D map are shown in the same directions as the projections above. Scale bar, 5 nm.

(E) Image reconstruction of the Gcd10/Gcd14 complex. C2 symmetry was imposed during the final rounds of refinement. The complex is shown along the symmetry axis (left) and perpendicular to the symmetry axis (right). Scale bar, 5 nm.

a soluble form; 86% success rate). Similarly, we were able to obtain soluble complexes corresponding to most of our validated targets using either complex reassembly or coexpression via either cotransformation of plasmids or single plasmids that contain operons encoding all of the proteins of interest (cf. Table 1; Supplemental Experimental Procedures) (9 of 11 complexes could be formed; 82% success rate). We believe that this achievement is principally due to the efficient selection criteria that we had established. It has been reported that only about 20% of full-length eukaryotic proteins are soluble when produced in a heterologous expression system (Graslund et al., 2008), but the performance of our approach is considerably superior. This is likely to be because only natively soluble proteins and complexes that are expressed at suitably high levels are detected by mass spectrometry after TAP purification, thereby biasing complex identification data toward soluble proteins.

Based on the 4 year experience of a consortium of numerous structural biology groups involved in 3D repertoire, we suggest an optimal experimental strategy for the high-throughput study of protein complexes. We conclude that, despite the absence of a "silver bullet," much can be achieved first by triaging the targets by an efficient computational procedure, followed by simple expression and reconstitution in the first instance. For this, a LIC-based strategy to clone optimized synthetic genes in a parallel manner resulted in notable success, with 14 of 16 subunits expressed in soluble form. During complex reconstitution, we had greater success when employing cosonication of *E. coli* in which each subunit had been expressed separately, compared with reconstitution using pure proteins and this has become our method of choice to obtain soluble complexes (cf. Supplemental Experimental Procedures: "Complex formation trials").

We also found that producing plasmids that encode the necessary subunits as synthetic DNA, with Shine-Dalgarno sequences upstream of the successive ORFs to be a very practical and rapid method of coexpressing complexes (cf. Supplemental Experimental Procedures: "Cloning strategy used for poly-cistronic expression"). Our studies into the use of polycistronic vectors, particularly those constructed from synthetic genes (e.g., Gcd10:Gcd14 and Ssl2:Yor352 complexes) (Figure S7) indicate that this is a strategy that this is a useful addition to pipelines, both because of the ease of production of plasmid constructs, and the increase in yield presented by codon-optimized genes.

In summary, we conclude that when initiating projects involving high-throughput study of protein complexes proper triaging and validation is obligatory. Despite clear advances in bioinformatics procedures, the direct inspection of the experimental primary data indicating the presence of a robust complex

by the stoichiometric interaction of its constituent components, as in our gel analysis of TAP-tagged complexes, remains the most effective method of complex selection. Once triaging and validation had been performed, it was relatively straightforward to test the association of the recombinant proteins experimentally. As we illustrate with the Gcd10:Gcd14 complex, we were able to obtain structural information during the relatively short timescale of this project. In this work, we have leveraged complementary strategies to the end of complex production for structural analysis, but we envisage the incorporation of further techniques in subsequent experiments. For example, high throughput small angle X-ray scattering studies of single proteins could be applied similarly to complexes (Hura et al., 2009), and it will be increasingly important to identify complex and subcomplex composition of samples purified directly from cells using native mass spectrometry (Hernandez et al., 2006). Accurate subunit prediction and validation methods will be beneficial to future high-throughput approaches geared toward "high-hanging fruit" and increase the probability that such efforts will yield illuminating insights into macromolecular machines at work.

EXPERIMENTAL PROCEDURES

Validation

TAP Purification

TAP-tagged strains of *Saccharomyces cerevisiae* were grown in 4 liters of YPD medium (1% yeast extract, 1% bacto-peptone, 2% glucose) to an optical density (O.D.) of approximately 2. Yeast pellets were resuspended in 40 ml of lysis buffer (1 mM DTT, 40 mM HEPES [pH 8], 250 mM NaCl) and frozen in liquid nitrogen. Cells were broken in a laboratory blender cooled with dry ice. Extracts were defrosted with protease inhibitors and spun in 35Ti rotor (Beckman) in a Beckman ultracentrifuge at 20,000 rpm for 20 min at 4°C. Supernatant was spun again at 32,000 rpm for 90 min at 4°C. Resulting extracts were dialyzed in buffer D (1 mM DTT, 40 mM HEPES [pH 8], 150 mM NaCl, 1 mM PMSF) and frozen in liquid nitrogen. Extracts were then defrosted and incubated with 200 μl of IgG Sepharose 6 Fast Flow resin (GE Healthcare) in the presence of 0.1% rTX-100 for 1.5 hr at 4°C. The beads were washed twice with 10 ml IPP150 (10 mM Tris-HCl [pH 8.0], 150 mM NaCl, 0.1% rTX100) and twice with 10 ml TEV cleavage buffer (10 mM Tris-HCl [pH 8.0], 150 mM NaCl, 0.5 mM EDTA, 1 mM DTT). TEV cleavage was performed for 2 hr using 20 μg of TEV protease in 300 μl of cleavage buffer at room temperature. Eluates were agitated with 300 μl of calmodulin beads suspension (Stratagene) for 0.5 hr at 4°C. The beads were washed four times with 500 μl of calmodulin wash buffer (10 mM Tris-HCl [pH 8.0], 150 mM NaCl, 10 mM β-mercaptoethanol, 1 mM $CaCl_2$) and the protein was eluted with 0.6 ml calmodulin elution buffer (10 mM Tris-HCl [pH 8.0], 500 mM NaCl, 10 mM β-mercaptoethanol, 0.1% rTX100, 4 mM EDTA). As a control, denatured elution fractions from both IgG and calmodulin beads were prepared with 250 μl of 1% SDS at 60°C.

Protein Precipitation and Analysis by Mass Spectrometry

Proteins were precipitated using pyrogallol red (Aguilar et al., 1999). When salinity of buffer was higher then 200 mM of NaCl, the samples were first adjusted to this concentration by dilution. Proteins were separated by electrophoresis performed on NuPAGE 4%–12% gradient gels using MES buffer gel system (Invitrogen) and stained with SimplyBlue SafeStain (Invitrogen). Mass spectrometry was performed both with IgG eluates in solution and from bands cut from gels. Samples were then processed by standard procedures with trypsin digestion and cysteine alkylation. The obtained peptide mixtures were separated on a nano-HPLC system and the column outlet was coupled to the ion source of an LTQ FTICR spectrometer.

Western Blot Analyses

After dialysis, extracts and flow-throughs after IgG Sepharose chromatography were separated by 10% SDS-PAGE and electro-blotted onto the Protran nitrocellulose membrane (Bioscience) using a Trans-Blot system (Bio-Rad). The filters were blocked for 1 hr in 5% milk powder in PBS containing 0.1% Tween 20 and then the mouse monoclonal anti-rabbit immunoglobulin-peroxidase conjugate (Sigma) diluted 3000-fold was added. After 1 hr, the blots were washed three times in PBS with 0.1% Tween 20. Finally, horseradish peroxidase conjugates were visualized by enhanced chemiluminescence system (ECL, GE Healthcare).

Mass Determination of the Complexes

In order to estimate the size of the purified complex, the extract from TAP-tagged strains was separated according to size, by size exclusion chromatography on a Superdex 200 10/300 column (GE Healthcare) using an Akta Purifier FPLC. Two different salt concentrations (150 and 500 mM NaCl) were used for elution and fractions were collected into a 96 well plate. Sixty microliters of every fraction was spotted on a nitrocellulose membrane. TAP-tagged subunits were detected by Dot-Blot as described for western blot analyses. The intensities of the spots were calculated with ImageQuant (GE Healthcare) and exported into chromatograms. The column was calibrated using protein markers; thyroglobulin (670 kDa), ferritin (440 kDa), catalase (232 kDa), aldolase (154 kDa), albumin (67 kDa), ovalbumin (43 kDa), and chymotrypsin (25 kDa).

Electron Microscopy and Image Processing

The purified, overexpressed Gcd10/Gcd14 complex was fixed on a glycerol gradient with glutaraldehyde according to the GraFix protocol (Kästner et al., 2008). Fractions of the gradient were further analyzed by dot-blot analysis using an antibody against the 6-histidine tag. The dot blot identified a single peak with a maximum at fraction 14. Samples from the peak fractions were prepared for subsequent electron microscopy by sandwich negative stain using uranyl acetate as previously described (Ulbrich et al., 2009). Samples were imaged at room temperature in a Philips CM120 Biotwin electron microscope at 100 kV. Data was recorded on a 4kx4k Tietz-CCD camera at a nominal pixel size of 4.27 Å per pixel under low-dose conditions. For further processing 10819 particle images were selected from 29 micrographs. Three-dimensional models were calculated using sinogram correlation and weighted back projection with IMAGIC 5 (van Heel et al., 1996). The process of determining initial orientations followed by calculation of a three-dimensional map was repeated several times using different class averages for starting the sinogram correlation.

Projections of the resulting three-dimensional models were compared with the initial class averages. The model that generated projections that matched 90% of the initial class averages was selected for further refinement by an iterative process of projection matching followed by calculating a new 3D map with Spider (Frank et al., 1996). After five rounds of refinement the map was stable and showed an approximately 4-fold symmetric core with extensions at opposite sides, giving the whole map a 2-fold symmetric appearance. Therefore, the map was refined for another five rounds imposing C2-symmetry. The final map is calculated from the best correlating 5503 particles. The resolution of the final map was determined by Fourier-Shell-Correlation and was 23 Å (correlation = 0.5) (Figure S8).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and four tables and can be found with this article online at doi:10.1016/j.str.2010.08.001.

ACKNOWLEDGMENTS

This work was supported in part by the European Union 6th Framework program "3D-repertoire" (LSHG-CT-2005-512028) to H.v.T., A.P., P.A., L.S. and to B.S. An EMBO young investigator award supported A.D. We thank Claire Batisse (EMBL Heidelberg) for technical assistance in electron microscopy. The authors declare that they have no conflict of financial interest with the work presented herein. L.S., P.A., and B.S. conceived of the study and M.B. coordinated the work between the contributing laboratories. P.A. and R.A.P. performed the bioinformatics analysis. K.G. and A.D. performed the in vivo validation. E.M. and M.C.-C. carried out the expression and coexpression testing using single plasmids and synthetic genes. M.A.B.,

N.L., and L.D. worked on expression and in vitro reconstitution of complexes. J.L., Y.B., M.C., V.H., D.L., and P.R. worked on complex production through various coexpression methods. N.L. and M.A.B. produced and characterized the Gcd10:Gcd14 and BB performed electron microscopy on it. B.S., H.v.T., A.P., and A.D. supervised the work in their laboratories. M.A.B. and A.P. wrote the article, with contributions from M.B., L.S., B.S., H.v.T., P.A., and A.D.

Received: February 21, 2010
Revised: June 30, 2010
Accepted: August 7, 2010
Published: September 7, 2010

REFERENCES

Aguilar, R.M., Bustamante, J.J., Hernandez, P.G., Martinez, A.O., and Haro, L.S. (1999). Precipitation of dilute chromatographic samples (ng/ml) containing interfering substances for SDS-PAGE. Anal. Biochem. *267*, 344–350.

Alzari, P.M., Berglund, H., Berrow, N.S., Blagova, E., Busso, D., Cambillau, C., Campanacci, V., Christodoulou, E., Eiler, S., Fogg, M.J., et al. (2006). Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. Acta Crystallogr. D Biol. Crystallogr. *62*, 1103–1113.

Anderson, J., Phan, L., Cuesta, R., Carlson, B.A., Pak, M., Asano, K., Bjork, G.R., Tamame, M., and Hinnebusch, A.G. (1998). The essential Gcd10p-Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA. Genes Dev. *12*, 3650–3662.

Anderson, J., Phan, L., and Hinnebusch, A.G. (2000). The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA *97*, 5173–5178.

Bravo, J., and Aloy, P. (2006). Target selection for complex structural genomics. Curr. Opin. Struct. Biol. *16*, 385–392.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., and Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet. *18*, 529–536.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., and Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. *116*, 190–199.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Graille, M., Chaillet, M., and van Tilbeurgh, H. (2008). Structure of yeast Dom34: a protein related to translation termination factor Erf1 and involved in No-Go decay. J. Biol. Chem. *283*, 7145–7154.

Graslund, S., Nordlund, P., Weigelt, J., Hallberg, B.M., Bray, J., Gileadi, O., Knapp, S., Oppermann, U., Arrowsmith, C., Hui, R., et al. (2008). Protein production and purification. Nat. Methods *5*, 135–146.

Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B., and Robinson, C.V. (2006). Subunit architecture of multimeric complexes isolated directly from cells. EMBO Rep. *7*, 605–610.

Higurashi, M., Ishida, T., and Kinoshita, K. (2008). Identification of transient hub proteins and the possible structural basis for their multiple interactions. Protein Sci. *17*, 72–78.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

Hura, G.L., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., 2nd, Tsutakawa, S.E., Jenney, F.E., Jr., Classen, S., Frankel, K.A., Hopkins, R.C., et al. (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat. Methods *6*, 606–612.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Kästner, B., Fischer, N., Golas, M.M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E., et al. (2008). GraFix: sample preparation for single-particle electron cryomicroscopy. Nat. Methods *5*, 53–55.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Marsden, R.L., and Orengo, C.A. (2008). Target selection for structural genomics: an overview. Methods Mol. Biol. *426*, 3–25.

Ozanick, S.G., Bujnicki, J.M., Sem, D.S., and Anderson, J.T. (2007). Conserved amino acids in each subunit of the heteroligomeric tRNA m1A58 Mtase from Saccharomyces cerevisiae contribute to tRNA binding. Nucleic Acids Res. *35*, 6808–6819.

Pache, R.A., and Aloy, P. (2008). Incorporating high-throughput proteomics experiments into structural biology pipelines: identification of the low-hanging fruits. Proteomics *8*, 1959–1964.

Shevchenko, A., Schaft, D., Roguev, A., Pijnappel, W.W., and Stewart, A.F. (2002). Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry: analytical perspective. Mol. Cell. Proteomics *1*, 204–212.

Smialowski, P., Martin-Galiano, A.J., Mikolajka, A., Girschick, T., Holak, T.A., and Frishman, D. (2007). Protein solubility: sequence based prediction and experimental verification. Bioinformatics *23*, 2536–2542.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. Science *320*, 1465–1470.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623–627.

Ulbrich, C., Diepholz, M., Bassler, J., Kressler, D., Pertschy, B., Galani, K., Bottcher, B., and Hurt, E. (2009). Mechanochemical removal of ribosome biogenesis factors from nascent 60S ribosomal subunits. Cell *138*, 911–922.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

